Language Acquisition and Learnability

Edited by
STEFANO BERTOLO

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

First published 2001

Printed in the United Kingdom at the University Press, Cambridge

Typeface *Times* System *3B2*

A catalogue record for this book is available from the British Library

ISBN 0 521 64149 7 hardback
ISBN 0 521 64620 0 paperback

Contents

1 A brief overview of learnability

Stefano Bertolo

Applications of formal learning theory to the problem of human language learning can be described as an exercise in which three parties – linguists, psychologists and learnability researchers – cooperatively construct a theory of human language learning and, in so doing, constrain their space of hypotheses by ruling out all the theories that violate one or more of the constraints that each party brings to bear on the problem.

The interaction among these three parties is similar to the interaction that would take place if a rich patron were to ask an architect and a structural engineer to work together to design a museum: the architect would start by designing very bold and innovative plans for the museum; the engineer would remind him or her, calculator in hand, that some of those designs would be physically impossible to build and the patron would visit every so often to make sure that the plans the engineer and the architect have agreed upon would result in a museum that could be built within budget and according to a specified construction schedule. In our case, linguists would correspond to the architect: based on their study of human languages or on more speculative reasons, they specify what they take the possible range of variation among human languages to be. Psychologists would correspond to the patron: they collect experimental data to show that it is not just that humans learn the language(s) of the linguistic community in which they are brought up, but that they do so according to a typical time schedule and relying on linguistic data of a certain, restricted, kind. Finally, learnability researchers correspond to the engineer: some theories of language variation they would be able to rule out directly, by showing that no conceivable mechanism could single out a correct hypothesis from such a large and dense range of choice; some other theories they would pronounce tenable, but only under certain assumptions on the resources available for learning, assumptions that need to be empirically validated by work in developmental psycholinguistics.

The goal of this introductory chapter is to provide linguists subscribing to Chomsky's *Principles and Parameters Hypothesis* (PPH) with a general understanding of the learnability concepts that need to be digested in order to study with profit research work at the intersection of linguistics, psychology and learnability and so, in particular, to provide the background that is required to understand the remaining essays in this collection.

1.1 The five components of a learning problem

Just like a structural engineer could not even *begin* to perform an analysis until a plan of the building has been provided together with the properties of the materials to be used in building it, a learnability researcher cannot even *begin* to work alongside linguists and psychologists until certain general properties of the learning problem to be solved are known:[1]

(i) What is being learned, exactly?
(ii) What kind of hypotheses is the learner capable of entertaining?
(iii) How are the data from the target language presented to the learner?
(iv) What are the restrictions that govern how the learner updates her conjectures in response to the data?
(v) Under what conditions, exactly, do we say that a learner has been successful in the language learning task?

We will briefly look at each of them in turn.

1.1.1 The end state of language learning

Since we are dealing with human language learning, the end state of this process is, by definition, a human language. In this section we will see, however, two things: that this fact itself has a number of interesting consequences and that there is disagreement of a rather interesting kind on what counts as knowledge of a human language.

First of all, since humans understand and produce utterances *productively* – i.e. they are able to understand and produce sentences they were never exposed to – having learned a human language cannot be equated with having memorized the list of all the sentences that one has ever encountered, but it must amount instead to having internalized a system of rules (a grammar). Under this view, the final state of the learning process encodes grammatical knowledge that can be used to classify every possible sentence as grammatical or ungrammati-

cal in the target language. This observation might give the impression that learnability may only address the problem of learning the surface syntax of a language. This is in fact not the case. One could easily annotate an utterance with all the relevant syntactic and semantic information. As long as this annotation results in a finite object that can be the input/output of a computation (see Wexler and Culicover (1980) for an example of how such an annotation can be carried out), the learning problem of finding a grammar for the given data remains essentially the same. It is for this reason that often in formal work learning problems are cast as problems of learning sets of natural numbers, the assumption being that appropriate coding could turn any learning problem into such a problem.

It is important to note that there is a fecund research tradition dating back to Horning (1969) that takes a different view of what the end state of the learning process is. Researchers in the field of stochastic grammars would claim that a grammar that explains the data and can be used productively is only part of what humans learn when they learn a natural language. In addition human learners also learn a probability distribution describing the applicability of the rules in the grammar. To exemplify, a learner would not only learn that two rules *A* and *B* are part of her grammar but also that, say, *A* is twice as likely to be used than *B* is. It is easy to see that the probability of every surface sentence in a language can thus be obtained by first determining how many possible structural interpretations the sentence can have and then adding all the values obtained by multiplying together the probabilities of each of the rules recruited in each interpretation. In this view, therefore, learners do not simply try to identify the grammar of their linguistic community but, rather, they try to approximate the ambient probability distribution on possible linguistic events.

All the contributions in this book proceed under the assumptions corresponding to the first view, but the reader must be aware of the fact that alternative views exist that are possibly better placed to explain certain facts about human language learning.

1.1.2 Available hypotheses

One of the idealizations on which much work on learnability relies is that learners must entertain hypotheses about the language they are trying to learn at every step of the way, so that, in effect, their learning history can be viewed as a data driven trajectory in a space of hypotheses, with the last state hopefully being (one of) the correct hypothesis(es).

For a learner to be successful, this space must contain at least a correct conjecture for each of the possible targets (if, while trying to learn Hungarian, you were not allowed to hypothesize a grammar equivalent to the grammar for Hungarian, you would naturally fail), but can otherwise be limited in many other ways. For example, it has been conjectured that, because humans converge on *productive* hypotheses about their target languages, it is perhaps a feature of their learning psychology that they cannot hypothesize any grammar that allows only finitely many sentences. The PPH has a very strong impact on this component of the learning problem. Its central idea is that human languages differ from each other only in finitely many respects (the *parameters*) and, in these respects, only in finitely many ways (the *values* of the parameters). But if this is so, all the hypothesis space needs to include are all the possible combinations of parameter values.

1.1.3 Learning environments

Generally speaking, given a possible sentence *s*, there are three kinds of clues the learner can receive about it from the environment: he or she can either be told, correctly or incorrectly, that *s* is part of her target language; or he or she can be told, correctly or incorrectly, that *s* is not part of her target language; or, finally he or she may not be told anything at all about *s*. In other words the environment, even when accurate, may provide less than complete information about the language to be learned. Martin Atkinson will show in his chapter that children learn language using evidence that comes from a rather constrained subset of the target language.

Finally, a common assumption in learnability is that learners are not aware of the rule – if any – according to which the environment is presenting the data. Why knowing such a rule would help for learning in the form of indirect negative evidence will be explained by the example on page 24 in the next chapter.

1.1.4 Learners

A finite sequence of sentences from the target language can be seen as an *evidential state* a learner could be in. Accordingly, learners can be broadly characterized by how they behave as a function of their present conjecture and evidential state.

For example, some learners base their next move in hypothesis space on the whole content of their evidential state (they have *perfect memory* of past data) while others only remember parts of it. Some learners

change their conjecture only when it is incompatible with their evidential state, others are not so constrained. Some try to modify their conjectures as little as possible in order to fit their evidential state, others take "wild guesses"... For at least one of the criteria of success we are going to examine, *identification in the limit*, there exists an impressive body of work detailing how the class of learnable languages changes when one or more of these restrictions are imposed on learners. The interested reader can consult Jain et al. (1999). Here we just mention the fact that, as we will see below when discussing *identification by enumeration*, while it is often easy to point out which parts of a learning procedure are at odds with one's best guess about the resources available to human language learners, it is often quite difficult to give a general characterization of the learning procedures one is comfortable with as far as developmental psycholinguistics is concerned.

1.1.5 Criteria of success

When can we say that a language learning strategy is successful? We will consider here three alternative criteria of success that are all compatible with the implicit premise of all the essays in this collection, the premise, that is, that what is learned is a grammar with no attached information about the probability distribution of the sentences that can be generated by it. As we introduce them we will show that it is very easy to prove that any class of languages that can be generated by a class of grammars consistent with the PPH is learnable under each of the three criteria.

The point of this exercise, which will be carried out using each time the same learning function, *identification by enumeration*, is to show that *all* PPH-consistent classes of languages are trivially learnable under *all* of the best understood criteria of success unless some rather substantial restrictions are imposed on what kind of learners humans are. The whole history of the last fifteen years of interaction between parametric linguistics and learnability (the essays in this volume included) can then be understood as an attempt to flesh out, using whatever evidence is available from empirical work in linguistics or psychology, what would follow from the assumption that human learners do not learn by enumeration but by some other, possibly independently motivated, mechanism. As mentioned above, no general consensus of what those mechanisms ought to be has emerged in the last fifteen years. As a result, rather than establishing general results, most recent studies have confined themselves to the analysis of individual learning algorithms, that as a consequence appear to the (ever shrinking) interested public as distant and isolated points in a vast and otherwise uncharted design space.

With all this in mind, let us start with our first criterion of success, *identification in the limit* (Gold, 1967).

1.1.5.1 Identification in the limit

Gold defines a *text* for a language L as an *infinite* sequence of sentences such that every sentence of L appears at least once in it and no sentence not in L ever appears in it. Let's consider now a procedure that takes as input *finite* initial segments of a *text* for a language and returns a conjecture (a grammar) about the language it is observing. Such a procedure is said to *identify* a *text* for L if and only if, after presentation of finitely many initial segments of the text it stabilizes on a single conjecture and that conjecture generates exactly L. The procedure is said to *identify* an entire language L if it is able to identify every possible text for it and, finally, it is said to identify a class of languages $\mathcal{L}$ if it identifies every L in it.

Given these definitions, in order to find out whether a class of languages $\mathcal{L}$ is learnable under the criterion of *identification in the limit* we either need to show the existence of a learning procedure that would *identify* every text for every language in $\mathcal{L}$, or show that no such learning procedure can exist.

We will immediately show that if a class of languages $\mathcal{L}$ has been generated by a set of grammars consistent with the PPH, then such a learning procedure does indeed exist. The procedure in question is called *identification by enumeration* (IBE) and was first described in Gold's (1967) seminal paper. We will describe it in detail because the very same learning procedure will be later employed to prove that PPH-consistent classes of languages are also learnable under the other two criteria of success we are going to discuss.

This is how IBE works for a finite class of languages such as those that result from any theory consistent with the PPH: the learner starts by writing down an enumeration $G_1, G_2 \ldots G_n$ of all her possible conjectures. The enumeration must have the property that, if $k > j$, then either $L(G_k) = L(G_j)$ (the language generated by grammar/hypothesis G_k is the same language as that generated by grammar/hypothesis G_j) or there is at least a sentence in $L(G_k)$ that is not in $L(G_j)$. She then initializes her hypothesis H to G_1 and the set of data observed D to the empty set. After presentation of each sentence s_i the learner first determines whether s_i is part of $L(H)$, the language resulting from her current hypothesis H. If so, she adds s_i to D and waits for a new sentence. Otherwise she adds s_i to D and changes her current conjecture H to the first G_i in the enumeration such that D is a subset of $L(G_i)$.

Now, let's see why IBE works for every finite class of languages. Suppose the target is $L(G_t)$, with $1 \leq t \leq n$ and σ is a text for the target. All we need to show is that: (a) IBE will never abandon the conjecture G_t – or any conjecture equivalent to it – if it ever happens to entertain it, and (b) there is a finite initial segment of σ in response to which IBE will hypothesize G_t or something equivalent. (a) follows immediately from the definition of IBE. As for (b) we can prove it by establishing that: (c) IBE will never hypothesize a G_k with $k > t$; (d) it will abandon every incorrect conjecture G_j after finitely many strings from σ, and (e) it will only entertain finitely many incorrect conjectures prior to hypothesizing G_t or one of its equivalents. Now, (e) directly follows from the fact that the enumeration $G_1 \ldots G_n$ is finite and (c) from the definition of IBE. As for (d), we can reason by contradiction: suppose that, in response to the first m sentences $s_1 \ldots s_m$ of σ, the learner conjectured a G_j such that $j < t$ and $L(G_j) \neq L(G_t)$ and never changed her conjecture thereafter. By definition of IBE, all of $s_1 \ldots s_m$ must be members of $L(G_j)$. Also, from the definition of the enumeration $G_1 \ldots G_n$ and the assumption that $j < t$ it follows that there is a sentence s_t in $L(G_t)$ that is not in $L(G_j)$. Now, since by assumption σ is a text for $L(G_t)$, s_t must appear somewhere in σ. And since s_t cannot be one of the $s_1 \ldots s_m$ sentences that caused the learner to hypothesize G_j, it must appear after the learner has hypothesized G_j. So, when the learner encounters s_t she is forced to abandon her conjecture, by the definition of IBE and the assumption that s_t is not a member of $L(G_j)$. But this contradicts the initial hypothesis that the learner could stick to the hypothesis G_j forever and so the proof by contradiction is completed.

So we now know that IBE is all that is needed to learn (identify in the limit) any parametric class of languages. With this established, we must hasten to add that identification in the limit is far too idealized a criterion of success to be used to model human language learning.

First of all it assumes complete and perfectly reliable information about the target language. In the following chapter, Martin Atkinson will show in detail what one would have readily suspected: the environment in which children learn their target language is *noisy* and fairly restrictive in the kind of information it makes available.

Second, it places no bounds of any kind on the amount of data learners are allowed to use to converge on their target: all that matters is that they do so *in the limit*. What we really would like is a criterion that would require the target to be reached after exposure to a number of sentences of the same order of magnitude as the number of sentences children are normally exposed to.

Finally, identification in the limit requires identification to be *exact*. But, as the whole chapter by Ian Roberts will explain, it is natural to argue that languages change over time precisely because there are situations in which learners get very close to the language of the previous generation, without, however, quite identifying it.

1.1.5.2 Wexler and Culicover's criterion

We'll take the cue from this last observation to introduce a second criterion of success on which a fair amount of research in linguistics and learnability depends. We will call it the Wexler and Culicover criterion, as it was most extensively used in Wexler and Culicover's (1980) seminal study on the learnability of transformational grammars. In more recent times it has been used as the criterion of choice in Gibson and Wexler's (1994) study and in all the research papers that extended their original idea, including the chapter by Sakas and Fodor in the present book.

While in identification in the limit it was possible for a sentence to appear exactly *once* in a text for a language, here we will require that at every time step in the learning process, every sentence in the target language have a nonzero probability of being presented to the learner. At the same time, instead of requiring *exact* identification of the target language we will just require that, for every $0 < \epsilon < 1$, for every language L in the class to be learned and every presentation sequence satisfying the condition above, there must be a finite number n such that, after presentation of n strings, the learner is guaranteed to output a conjecture that has probability less than ϵ to be incorrect.

Now, it is very easy to show that IBE can also learn every finite class of languages (and so every class of languages generated by a PPH theory) under the Wexler and Culicover criterion. The proof goes like this: since every text satisfying the Wexler and Culicover criterion is also a text in Gold's sense, it follows that after finitely many strings IBE can identify exactly, that is with error $\epsilon = 0$, every text in the sense of Wexler and Culicover. But if it can do so for $\epsilon = 0$, it can do so for every $0 < \epsilon < 1$.

1.1.5.3 PAC learning

Still, even in the Wexler and Culicover criterion, there is no requirement that the number of sentences that are used to attain convergence with less-than-ϵ error be a *small* number for every language in the class to be learned. This concern is, on the contrary, at the heart of the last criterion we will consider here, *Probably Approximately Correct* (PAC) learning, a criterion first proposed by Valiant (1984).

According to the PAC criterion, a learner is successful if and only if she has a *very high probability* of producing a conjecture that is *very close* to the target language when using only *reasonably few* sentences from the target language.

The general idea underlying PAC learning is that data are presented to the learner according to a probability distribution unknown to her and different for each target language. In other words, every language has, associated with it, a function that, for each sentence in it, determines how likely the given sentence is to be presented to the learner at any one time.[2]

PAC captures the idea that learners are successful if and only if they rapidly produce conjectures that would misclassify only a set of sentences from the target language that has a negligible chance of appearing in the learner's environment. How rapidly they are required to produce such a conjecture is a function of how negligible the chance of error is required to be: the smaller the chance of error required, the larger the size of the sample that can be used. Larger samples, naturally, take longer to be collected.

Moreover, unlike the Wexler and Culicover criterion, PAC does not require that a less-than-ϵ-error conjecture be *always* arrived at, but only that the probability of it not being arrived at be smaller than a certain confidence parameter δ ($0 \leq \delta < 1/2$).

However, as mentioned before, PAC is rather strict on the size of the sample that can be consumed to produce a conjecture that has less than δ probability of being in error by more than ϵ: it requires that, for every choice of δ and ϵ, the size of the sample consumed be a polynomial function of $1/\delta$ and$1/\epsilon$. In other words, the size of the sample should not grow *very fast* as $1/\delta$ and $1/\epsilon$ grow.

Once again, IBE can be used to show that every finite class of languages is PAC learnable. This follows directly from a theorem proved by Blumer et al. (1987) which states that a class of languages is PAC learnable if an *Occam learning algorithm* can be found for it.

Blumer defines an Occam learning algorithm for a class of languages $\mathcal{L}$ as an algorithm that, in response to data from any L in $\mathcal{L}$ always outputs a conjecture that has two properties, it is consistent with the sample received and has a complexity not exceeding a certain value that is a function of the least complex hypothesis available for the language to be learned and the size of the sample itself. The interested reader can find the precise statement of this second requirement in Blumer's original article. For the purpose of the present discussion, it is sufficient to note that the requirement in question is designed to disqualify hypotheses that, even if consistent with the data, consist

simply of an enumeration of the sample itself. It will soon be clear, however, that the details of this requirement are not important for our proof.

So, we want to show that IBE is an Occam learning algorithm for every finite class of languages (when the space of hypotheses is equally finite). That its conjectures are always consistent with the sample presented follows immediately from the definition of IBE. As for the second requirement, it is sufficient to observe that since the learner can entertain only finitely many hypotheses, there exists an upper bound on the complexity of the possible conjectures. The reader may consult Robin Clark's chapter in this book for a tutorial on how the complexity of hypotheses may be defined. Here we just note that this upper bound is a constant and that this is sufficient to meet the second part of Blumer's requirement for Occam learning algorithms. The reader can refer to Blumer's original paper to be convinced that this is indeed the case. This completes the proof that every finite class of languages, and so any class that is consistent with the PPH, is PAC learnable.

1.2 Moving away from identification by enumeration

Our insistence on using IBE in the proof that PPH-informed classes of languages are learnable under each of the success criteria we examined is due to the importance of driving home the following point: proving the learnability of PPH-informed classes of languages is generally speaking a trivial enterprise, whichever criterion of success one decides to adopt. The task becomes challenging only once one decides to reject as part of a model of human language learning the two very features that make IBE as successful as it is: complete memory of all past data and access to an initial enumeration that allows for hypotheses to be searched in the *correct order*.

Even in the absence of much substantive work in developmental psycholinguistics showing how much memory for past data children can be expected to have in the process of language learning, it is quite reasonable to reject the hypothesis that they have recall of their entire learning sample. It is interesting to note that most recent models of parametric language acquisition such as Dresher and Kaye's cue based learner, Gibson and Wexler's TLA, Clark's Genetic Algorithm and Fodor's STL (see Martin Atkinson's and Fodor and Sakas' chapters for detailed discussion of each of them) make the exact opposite assumption and attempt to show how learnability may be proved even using learners that have no explicit memory whatsoever of their learning sample although some of them cleverly encode some of it in the conjectures they entertain at any given time.

But it is when we start exploring what it would mean to abandon the *other* important feature of IBE, the prescient enumeration of all possible hypotheses, that we finally get to the heart of what makes the interaction between parametric linguistics, psychology and learnability interesting. IBE was repeatedly proved to work under the assumption that only finitely many hypotheses are available to the learner. But the PPH is a much stronger hypothesis than that: it not only implies that the class of possible (hypotheses about) human languages is finite but, most importantly, that variation among human languages can be expected to be systematic and take place along several, largely orthogonal dimensions. What this suggests, in effect, is that, although it is unrealistic to expect that human learners rely on a space of hypotheses that is nicely linearized as seen in IBE, it is perhaps not unreasonable to suppose that it is organized according to other principles that may make it possible to search it reliably and efficiently. The last fifteen years of work on the subject could be fairly reconstructed as a sustained investigation of what those principles might be. The entire line of research on the so called *Subset Principle* initiated by Wexler and Manzini (1987) and summarized by Martin Atkinson in the next chapter can be seen as an investigation of the structure of the space of hypotheses that would result if all parameter values except one were kept constant.

More recent work has concentrated on the possible patterns of overlapping among target languages that result from alternative settings of the parameters and allow algorithms as diverse as Gibson and Wexler's TLA and Fodor's STL to search the resulting space of hypotheses reliably and efficiently.

Because such patterns, which will form the subject of the discussion in much of the rest of this book, are often quite intricate, this introductory chapter will end by introducing a formal definition of parameter spaces that will, in effect, serve as a specialized vocabulary designed to make it possible to describe those patterns precisely and concisely. All the contributors to this book will then be able to use this vocabulary as a *lingua franca* that will facilitate the understanding of several, interrelated arguments by translating them from their original formulation into a common standard.

1.3 A formal model of parameter spaces

In this final section we will proceed as follows: we will first give a definition of parameter spaces that is general enough to cover any conceivable theory of language variation consistent with the PPH. We will

then introduce the notation that is necessary to isolate regions of interest in any parameter space of interest.

Definition 1.1 *A parameter space $\mathcal{P}$ is a triple $\langle par, L, \Sigma \rangle$, where Σ is a finite alphabet of symbols and par is a finite set of parameters $\{p_1 \ldots p_n\}$. The parameters themselves can be seen as sets. In particular, given a parameter p_i, its members, which we enumerate as $v_i^1 \ldots v_i^{|p_i|}$, are just the "possible values" of p_i. Given the cartesian product $\mathbf{P} = p_1 \times p_2 \times \cdots \times p_n$, a parameter vector $\overline{P}$ is a member of $\mathbf{P}$, namely a possible way of choosing one value for every parameter. Let Σ^* be the set of all finite strings over the alphabet Σ and 2^{σ^*} its power set (the set of all subsets of Σ^*). The function $L : \mathbf{P} \mapsto 2^{\Sigma^*}$ assigns a possibly empty subset of Σ^* to each vector $\overline{P} \in \mathbf{P}$, that is it associates every possible parameter setting (vector) with a language (set of strings).*

As it turns out, in order to analyze a parametric learning algorithm it is important to have a way of referring to portions of a parameter space that share a certain value assignment to certain selected parameters. This can be achieved in two steps by first defining the notion of a *partial* parameter assignment and then showing how a parameter space can be *sliced* according to a particular parameter assignment.

Definition 1.2 *Let $\mathcal{P}$ be a parameter space. A partial assignment in $\mathcal{P}$ is any subset B of*

$$\bigcup_{p_i \in par} \{p_i\} \times p_i$$

such that for every p_i in par there is at most one $\langle p_i, v_i^m \rangle$ in B. Given two partial assignments A and B in $\mathcal{P}$, B is said to be A-consistent iff $A \cup B$ is also a partial assignment in $\mathcal{P}$.

For example, suppose we have a parameter space with two binary valued parameters p_1 and p_2. If $p_1 = \{0, 1\}$, then the cartesian product $\{p_1\} \times p_1$ is the set of all ordered pairs such that the first element is the parameter p_1 and the second element is one of its values, 0 or 1. The cartesian product $\{p_1\} \times p_1$ therefore turns out to be $\{\langle p_1, 0\rangle, \langle p_1, 1\rangle\}$. Each of the elements of the cartesian product can be seen as a particular assignment of value to p_1. If p_2 is also equal to the set $\{0, 1\}$, then we have that

$$\bigcup_{p_i \in par} \{p_i\} \times p_i = \{\langle p_1, 0\rangle, \langle p_1, 1\rangle, \langle p_2, 0\rangle, \langle p_2, 1\rangle\}.$$

By definition 1.2 the subsets $\{\langle p_1, 0\rangle, \langle p_2, 0\rangle\}$, $\{\langle p_1, 0\rangle\}$ and $\{\langle p_2, 0\rangle\}$ are all partial assignments in $\mathcal{P}$. In particular, $\{\langle p_2, 0\rangle\}$ is $\{\langle p_1, 0\rangle\}$-consistent.

Finally, the following definition formalizes the notion of concentrating only on a part of a parameter space, a part that is picked up by choosing a partial assignment. The function π_i is a "projection" function: it takes a vector (parameter assignment) and returns the i-th element of the vector.

Definition 1.3 *Let $\mathcal{P}$ be a parameter space and A a partial assignment in it. If $A = \varnothing$, then $\mathcal{P}[\mathcal{A}] = \mathcal{P}$. If $\mathcal{P}[A]$ is a parameter space $\langle par^A, L, \Sigma\rangle$ (with $par^A = \{p_1^A \ldots p_n^A\}$) and B is an A-consistent partial assignment in $\mathcal{P}$, then the subspace $\mathcal{P}[A \cup B]$ is the parameter space $\langle par^{A\cup B}, L, \Sigma\rangle$ such that, given*

$$H = \bigcup_{x\in B} \pi_1(x),$$

if $p_j \notin H$ then $p_j^{A\cup B} = p_j^A$ and if $p_j \in H$ then $p_j^{A\cup B} = \{v_j^m\}$ where v_j^m is the only $v \in p_j$ such that $\langle p_j, v_j^m\rangle \in B$. Finally, $\mathcal{P}[A \cup \overline{B}]$ is the parameter space $\langle par^{A\cup\overline{B}}, L, \Sigma\rangle$ where, for every p_i in H, $p_i^{A\cup\overline{B}} = p_i^A - p_i^{A\cup B}$ and, for every p_i not in H, $p_i^{A\cup\overline{B}} = p_i^A$.

This definition looks more formidable than it really is. All it says is that, given a parameter space $\mathcal{P}$, and a partial assignment A, it is possible to define the "$\mathcal{P}[A]$ corner" of $\mathcal{P}$ to consist of all and only those languages in $\mathcal{P}$ that have their parameters set exactly as dictated by A. So, if $A = \{\langle p_i, 0\rangle, \langle p_j, 1\rangle\}$, the "$\mathcal{P}[A]$ corner" of $\mathcal{P}$ contains all and only those languages that have p_i set to value 0 and p_j set to value 1. Similarly, $\mathcal{P}[\overline{A}]$ can be seen as the "mirror image" of the "$\mathcal{P}[A]$ corner" since it is the set of all languages that have all the parameters listed in A set differently than is dictated by A.

It is worth noting that the larger the partial assignment A, the smaller the "$\mathcal{P}[A]$ corner" of $\mathcal{P}$ one is looking at. This means that, when we take a partial assignment A and we expand it to a larger partial assignment $A \cup B$, the "$\mathcal{P}[A \cup B]$ corner" of $\mathcal{P}$ is smaller than the original "$\mathcal{P}[A]$ corner".[3]

This is all the descriptive apparatus that we will need for the discussion that follows. So, for example, using the notation

$$s \in \bigcup_{\overline{P}\in\mathbf{P}^A} L(\overline{P}) - \bigcap_{\overline{P}\in\mathbf{P}^{A\cup\overline{B}}} L(\overline{P})$$

we can avoid the cumbersome expression "s belongs to at least one of the languages whose parameters have been set as dictated by A but it is

not the case that it belongs to every language that has its parameters set as dictated by A and at variance with all the parameter assignments listed in B''. Hopefully, the reader will soon appreciate the advantages of the notation.

Notes

1 Sadly, much work on language learning is published in which entire theories are established and conclusions are drawn with disregard for these very basic requirements.
2 PAC is really able to represent a more general scenario, where what is assigned a probability is the event that a given sentence is presented to the learner as a positive *or* negative example of the target language.
3 This obviously assumes that B is not empty and that B is not properly included in A.